Statistical Methods for Learning Curves and Cost Analysis

Matthew Goldberg, CNA Corporation
Anduin Touw, Boeing Corporation

January 2003

Cost Analysis Statistical Monograph

- Written over the past four years with Anduin Touw (Boeing Corporation)
- Inspired by David Lee's "*The Cost Analyst's Companion*"
- Greater emphasis on statistical methods for learning curves and CERs
- To be published by INFORMS, *Topics in Operations Research* series, 2003

Questions I Have Been Asked As A Book Author

- How many pages is your book?

  180 pages

- OK, then, what’s it about?

  Statistical Methods for Learning Curves and Cost Analysis

- Really, may I have a (free) copy?

  No, but it will be priced very reasonably (well under $50 / copy, in paperback)

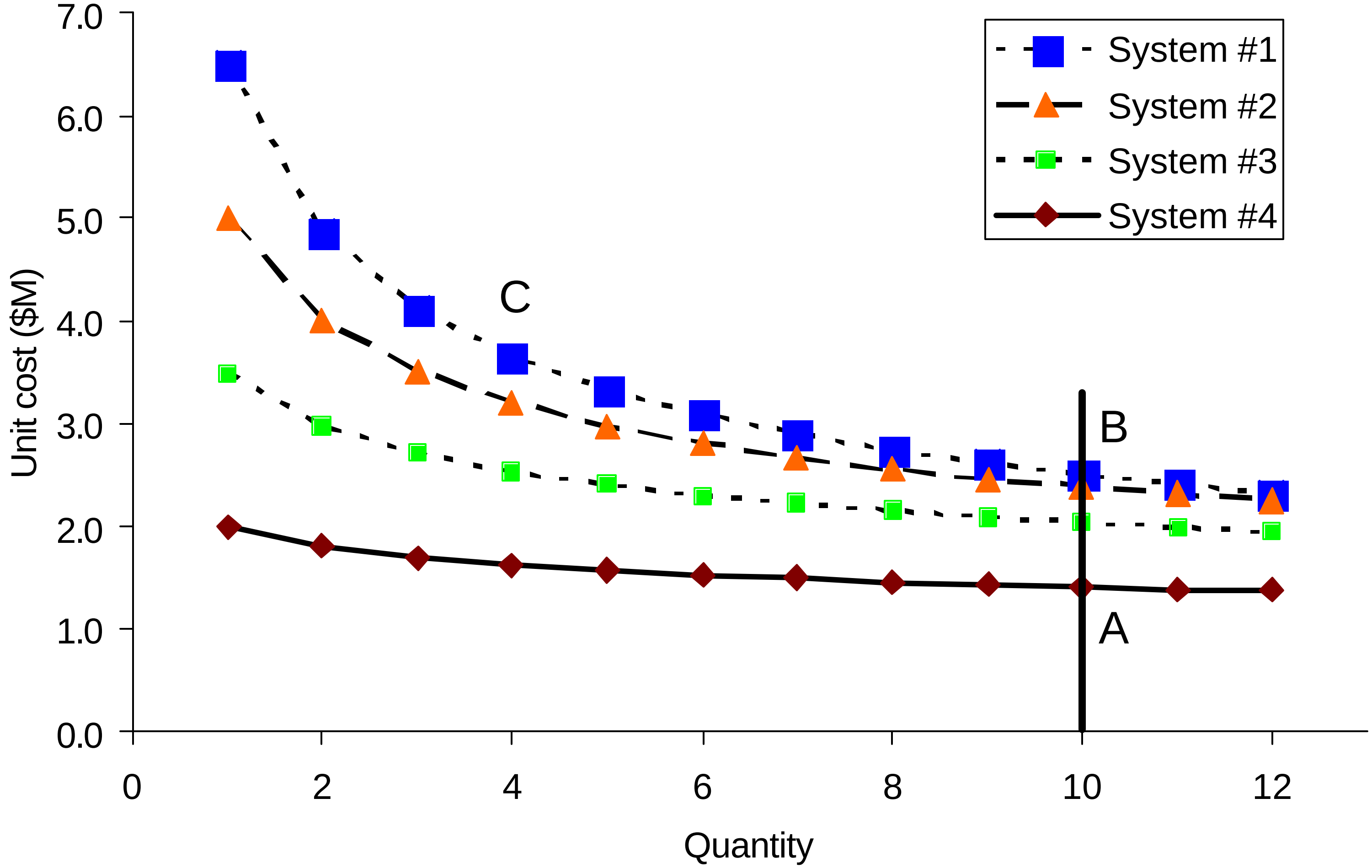
- Daddy, are you really smart enough to write a book?

  No, David, I’m not

Statistical Problems in Military Cost Analysis

- Marginal cost of a weapon system varies with:
  - technical and performance characteristics
    - e.g., weight, speed, materials content
  - unit number in the production sequence
    - “learning”
- Learning curve: relationship between cost and sequence number, holding fixed the technical and performance characteristics
- Cost estimating relationship (CER): relationship between cost and characteristics, holding fixed the sequence number (e.g., 100th unit)

Learning Curve versus CER: Two Perspectives



Data on Sequential Production Lots

- We do not typically observe data on individual production units
- Instead, we observe data on "lots"
  - typically annual lots, though a given lot may span several fiscal years start-to-finish

Lot number	Lot start	Lot end	Lot size	Incremental lot cost ($M)	Lot average cost ($M)
1	1	218	218	102.765	0.471
2	219	1,158	940	212.158	0.226
3	1,159	3,200	2,042	321.819	0.158
4	3,201	5,900	2,700	333.720	0.124
5	5,901	7,591	1,691	212.558	0.126
6	7,592	10,011	2,420	227.238	0.094
7	10,012	11,668	1,657	157.912	0.095
8	11,669	14,436	2,768	171.339	0.062

Learning Curve Estimation: Lot Midpoint Iteration

- Power-function model for marginal cost:

$$MC(Q) = T_1 \times Q^b, \quad -1 < b \le 0$$

- Incremental lot cost:

$$TC_i - TC_{i-1} = \int_{Q_{i-1}+0.5}^{Q_i+0.5} T_1 \, z^b dz = \frac{T_1}{1+b} \times \left[(Q_i + 0.5)^{1+b} - (Q_{i-1} + 0.5)^{1+b} \right]$$

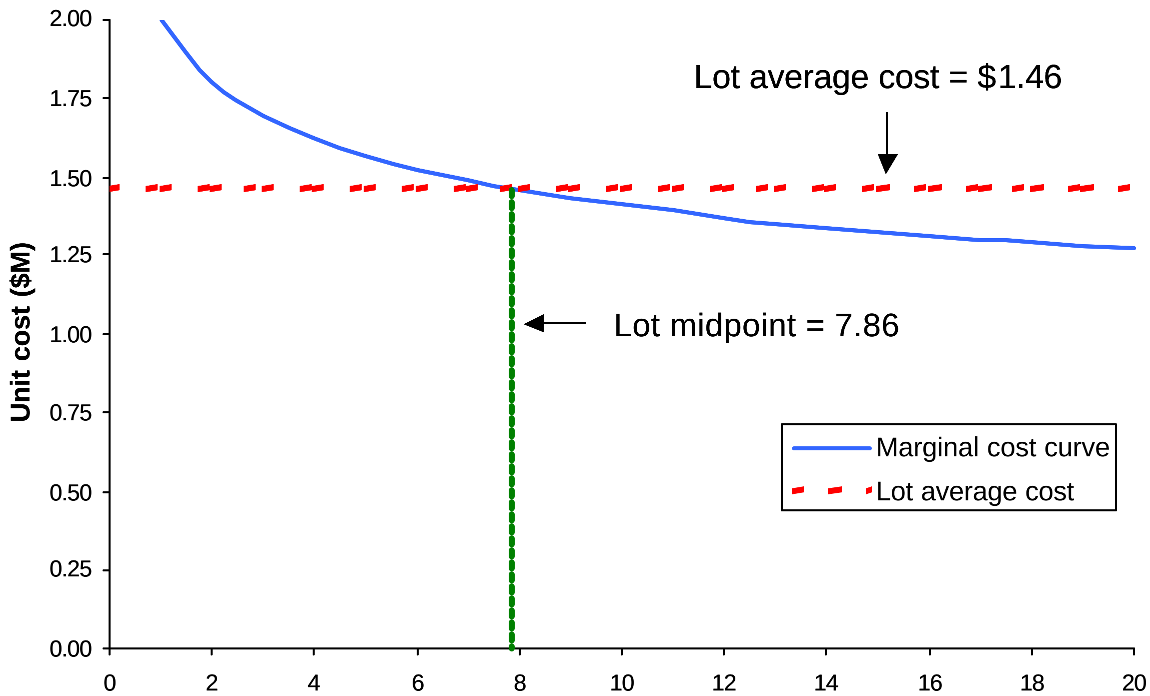
- Lot average cost:

$$LAC_i = \frac{TC_i - TC_{i-1}}{Q_i - Q_{i-1}} = \frac{T_1}{(1+b) \times (Q_i - Q_{i-1})} \times \left[(Q_i + 0.5)^{1+b} - (Q_{i-1} + 0.5)^{1+b} \right]$$

Lot Midpoint Calculation

- Find the point interior to each lot whose marginal cost equals the lot average cost

$$AC_i = MC\,[\overline{Q}_i(b)] = T_1 \times [\overline{Q}_i(b)]^b \longrightarrow \overline{Q}_i(b) = \left(\frac{[(Q_i + 0.5)^{1+b} - (Q_{i-1} + 0.5)^{1+b}]}{(1+b)\times(Q_i - Q_{i-1})} \right)^{1}$$



Lot Midpoint Iteration

- By the definition of the lot midpoint:

$$LAC_i = MC[\overline{Q}_i(b)] = T_1 \times [\overline{Q}_i(b)]^b, \quad i = 1, \ldots, n$$

- Take natural logarithms:

$$\ln(LAC_i) = \ln(T_1) + b \ln[\overline{Q}_i(b)], \quad i = 1, \ldots, n$$

- Alternate between these two steps, until (hopefully) convergence
  - Calculate the midpoint of each lot *i*
  - Run a linear regression on the midpoints, for lots $i = 1, \ldots, n$

Assessment of Lot Midpoint Iteration

- Lot-midpoint iteration has no theoretical foundation
  - There may not be a “root”
  - Root may not be unique
  - Iteration may not converge to any root
  - Does not optimize any continuous function
    - Maximize likelihood function
    - Minimize sum-of-squares

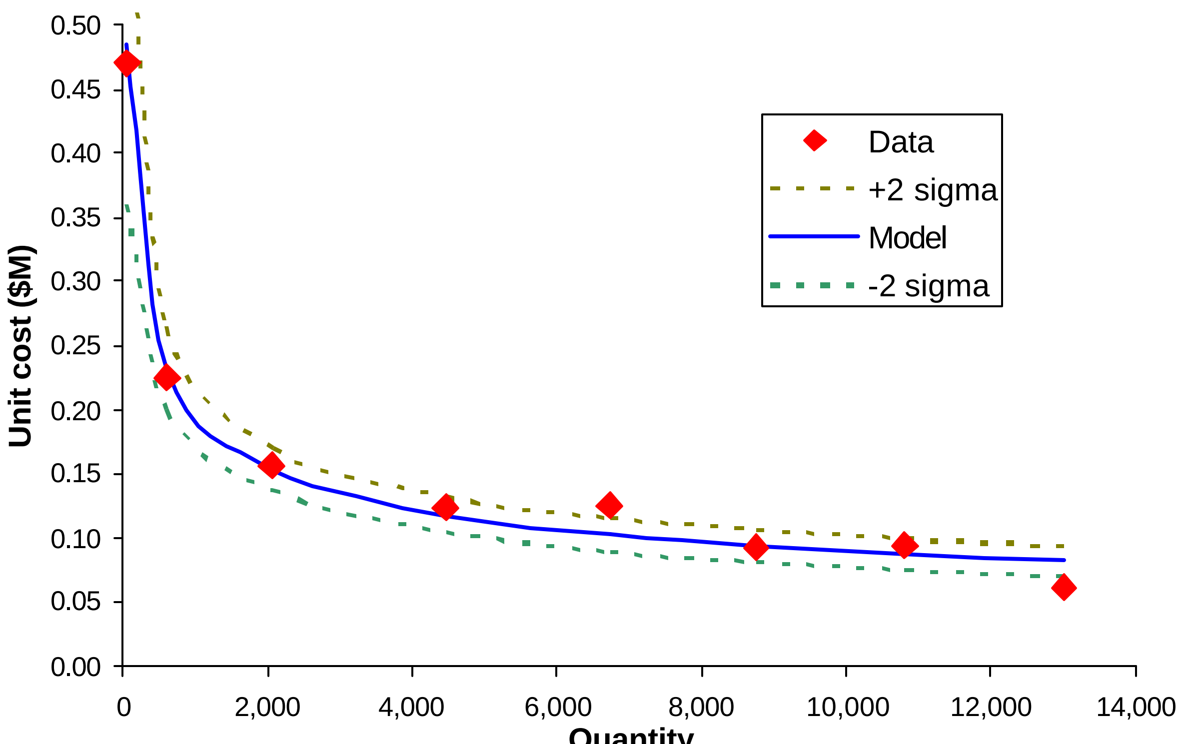
Convergence example

- Better to use non-linear least squares (NLS)

Confidence band for the NLS predictions, using lot midpoints as plot points

$$Var(L\hat{A}C_i) \;\rightarrow\; \left[w_i \, V \, w_i^T \;+\; \frac{s^{\,4}}{2n} \right] \times (L\hat{A}C_i)^{\,2} \,,$$

$$\text{where}\;\; w_i \;=\; \left(1\big/\hat{T}_1 , \quad \ln \bar{Q}_i \right)$$



Models with Multiplicative Error Structures

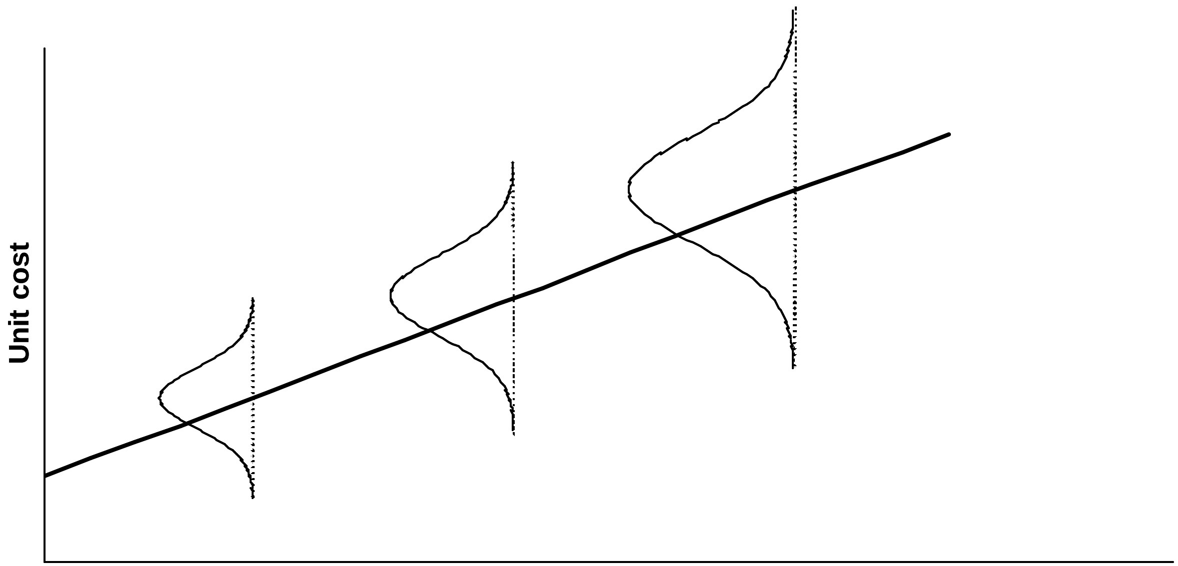
- The learning curve:

$$LAC_i = T_1 \times \overline{Q}_i^{\,b} \times (1+u_i)$$

or the CER:

$$\text{Unit cost}_i = b_0 \times \text{Weight}_i^{b_1} \times (1+u_i),$$

$$\text{where } Var(u_i) = s^2 \text{ for all } i = 1, \ldots, n$$



Minimum Percentage Error (MPE) Estimation

- Lee, Book have proposed to minimize the sum of percentage errors:

$$\hat{b} = \arg\min_{b} \sum_{i=1}^{n} \left(\frac{y_i - f(x_i, b)}{f(x_i, b)} \right)^2$$

  – Estimates are biased, even in very large samples
  – Estimates are sensitive to outliers

- Better to use iteratively reweighted least squares (IRLS)

Why MPE is Biased: Two Perspectives

- If error terms are multiplicative normal, then log-likelihood function looks like:

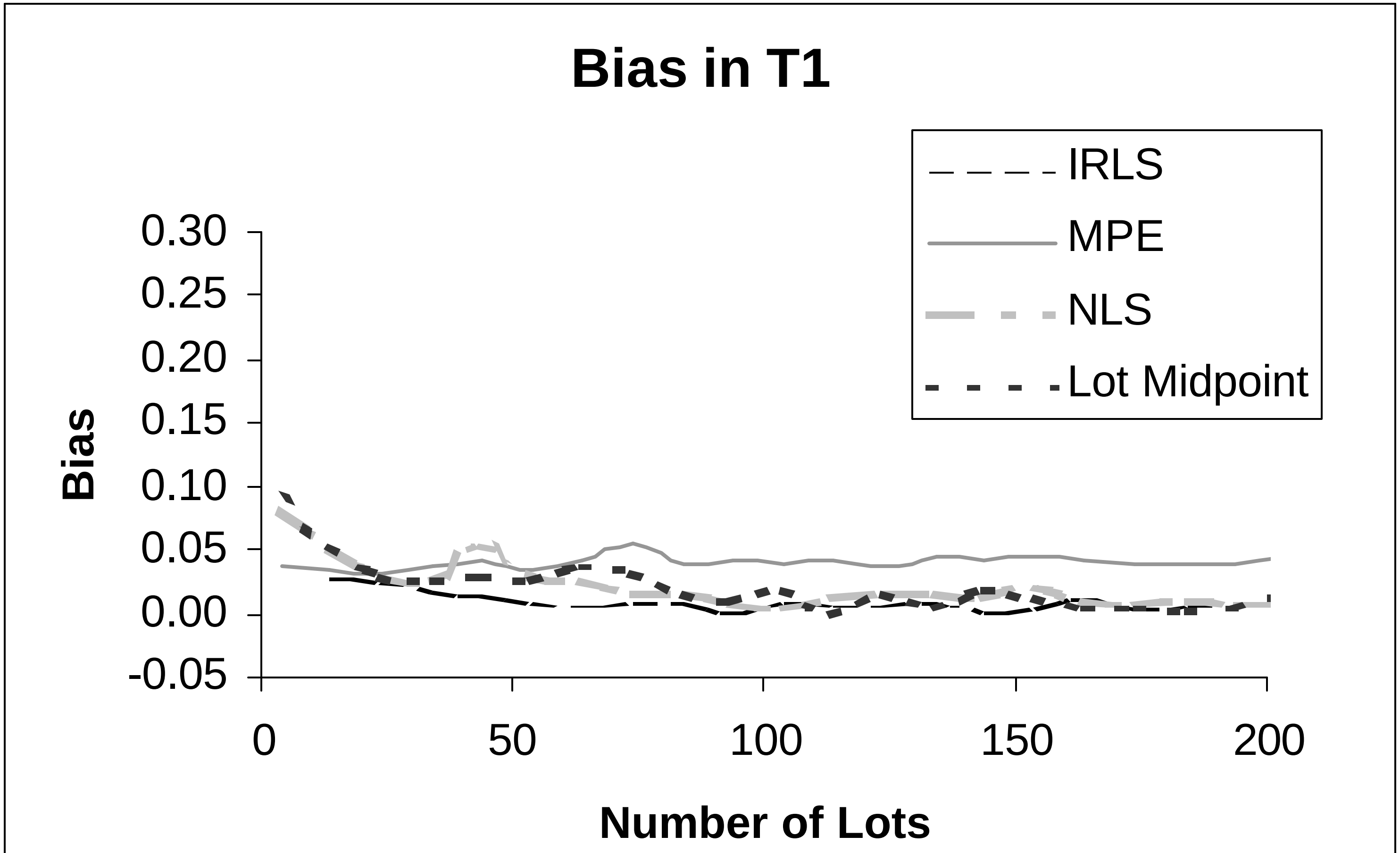
$$-\sum_{i=1}^{n}\left(\frac{y_i - f(x_i, b)}{f(x_i, b)}\right)^2 + \text{additional term involving } b$$

  – by dropping the additional term, you shift the location of the maximum away from unbiased MLE

- The minimization algorithm is “tempted” to minimize the sum of percentage errors by inflating the denominator
  – model predictions are biased high, particularly model intercept

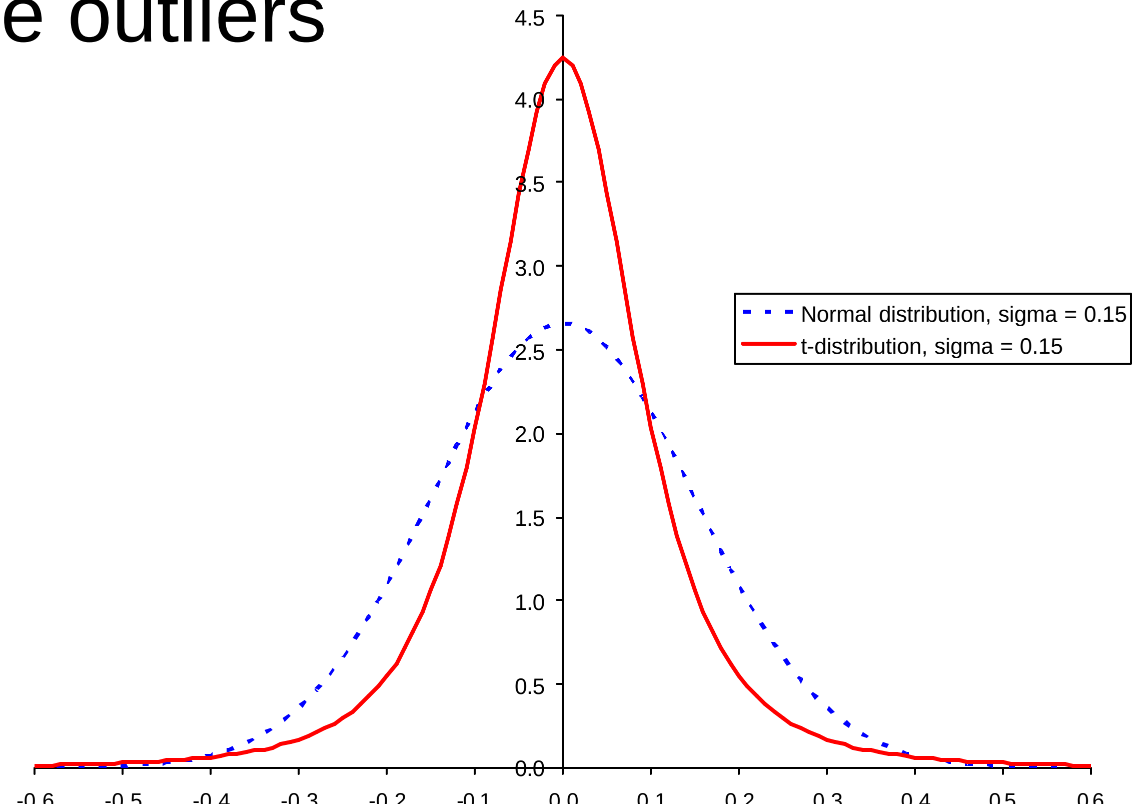
Monte Carlo Results with normally-distributed errors

Normal errors, $T_1 = 1.8$, $b = -0.33$ (80% slope), $\sigma = 0.15$



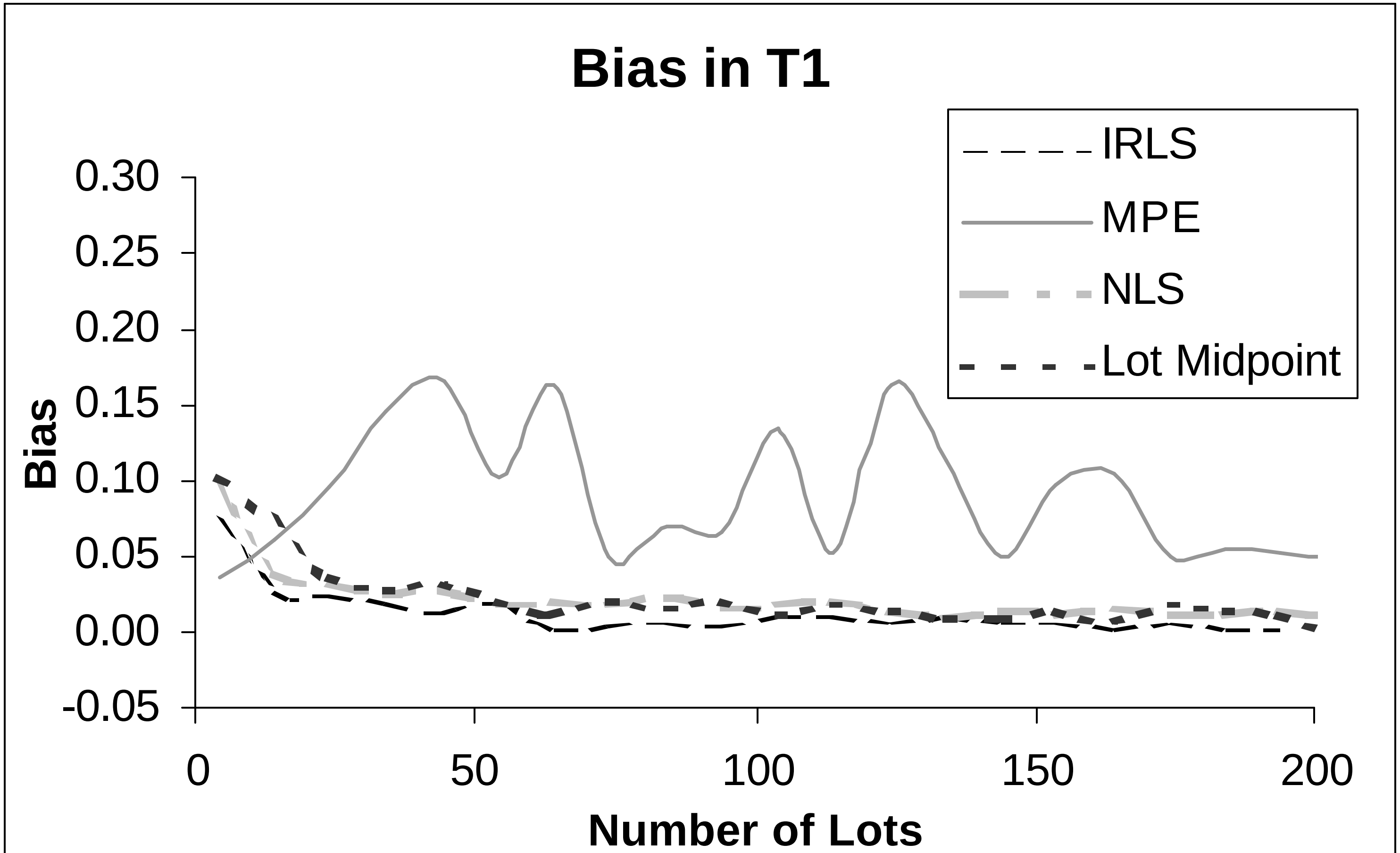
Sensitivity of Estimators to Outliers (*t*-distribution)

- *t*-distribution with 3 degrees-of-freedom, normalized to have $\sigma = 0.15$
- greater probability in the tails $\left(|u| > 2.80 \times s\right)$? more outliers



Monte Carlo Results with *t*-distributed errors

t-distributed errors, $T_1 = 1.8$, $b = -0.33$ (80% slope), $\sigma = 0.15$



Theoretical Comparison of Estimation Methods

Estimation Method	Distributional Assumptions	Asymptotic Properties	Software Availability	Covariance Matrix
Lot midpoint non-linear least squares (NLS)	multiplicative model, log-normal errors	consistent and asymptotically normal	any statistical package; or manually in Excel Solver (but no covariance matrix)	automatic in any statistical package; feasible in Excel
Lot-midpoint iteration	multiplicative model, log-normal errors	unknown	manually in Excel; programmable in SAS	conventional formula, an underestimate
Minimum percentage error (MPE)	multiplicative model, finite variance	biased and inconsistent	Excel Solver; programmable in SAS	unknown
Iteratively reweighted least squares (IRLS)	multiplicative model, finite variance	consistent and asymptotically normal	some statistical packages (e.g., SAS); manually in Excel Solver	automatic for supporting statistical packages; feasible in Excel

Backup slides

Convergence

Return

- Consider the sequence $b, b^2, b^3, \ldots$
- This sequence converges if $|b| \le 1$
  - e.g., $1/2, (1/2)^2, (1/2)^3, \ldots \quad \to 0$
  - or $-1/2, (-1/2)^2, (-1/2)^3, \ldots = -1/2, 1/4, -1/8, \ldots \quad \to 0$
- It diverges if $|b| > 1$
  - e.g., $2, 2^2, 2^3, \ldots \quad \to \infty$
- Non-linear, multi-variate generalization:
  iteration converges if all eigenvalues of Jacobian matrix <1 in absolute value